Pathfinder: Multiresolution Region-based Searching of Pathology Images using IRM *

James Z. Wang, M.S. Math, M.S.C.S.†

Stanford Medical Informatics and Computer Science
Stanford University, Stanford, CA 94305

The fast growth of digitized pathology slides has created great challenges in research on image database retrieval. The prevalent retrieval technique involves human-supplied text annotations to describe slide contents. These pathology images typically have very high resolution, making it difficult to search based on image content. In this paper, we present Pathfinder, an efficient multiresolution region-based searching system for high-resolution pathology image libraries. The system uses wavelets and the IRM (Integrated Region Matching) distance. Experiments with a database of 70,000 pathology image fragments have demonstrated high retrieval accuracy and high speed. The algorithm can be combined with our previously developed wavelet-based progressive pathology image transmission and browsing algorithm and is expandable for medical image databases.

INTRODUCTION

The World-Wide-Web (Web) has revolutionized medical education by allowing easy access to educational materials and providing new modes of interaction with these materials. We have developed a multiresolution region-based image retrieval technique to search and browse high-resolution pathology slides. Our goal is to develop a Web-based "smart virtual microscope" for image viewing that (1) allows users to locate quickly regions-of-interest in a large pathology slide database (2) minimizes the manual effort needed to build medical image libraries, and (3) reduces the network load.

Images captured from slide preparations are typically large and have very high resolution, making it difficult to search based on image content. Moreover, it is inefficient to transmit the full images over the network for educational purposes. The prevalent retrieval technique involves human-supplied text annotations describing slide contents. Searching is then performed based on keywords or full-text. This approach is easier to implement than content-based retrieval because of the support from the existing textual database technology. However, it is problematic for several reasons: it does not allow students to find regions-of-interest based on image features, such as shape and object configurations; it involves an enormous amount of manual effort to build image libraries; the image description may not be consistent with the user query; and it wastes a lot of network bandwidth to transmit the whole image.

Content-based image retrieval (CBIR) is defined as the retrieval of relevant images from an image database on the basis of automatically-derived imagery features. Many general purpose CBIR systems have been developed, such as the IBM QBIC System [3] developed at the IBM Almaden Research Center, the Photobook System developed by the MIT Media Lab [9], the WBIIS System [12] developed at Stanford University, the Blobworld System [1] developed at U.C. Berkeley, and the SIMPLIcity System [6, 14] developed at Stanford University. These systems can not be applied to pathology images or radiology images because of the special characteristics of these images: extremely high resolution and limited colors.

Figure 1: Automatic object segmentation of pathology images is an extremely difficult task. We avoid the precise object segmentation process by using our IRM "soft matching" metric.

Among others, researchers of Rutgers University have developed an CBIR system for pathology images [2] in 1998. The system is capable of searching pathology images based on shape. Experiments with

*This work was supported in part by the National Science Foundation under Grant No. IIS-9817511. Project URL: http://www-db.stanford.edu/IMAGE/

†Ph.D. expected in 2000. wangz@cs.stanford.edu

a database of 261 color images have shown promising results. However, the system relies on correct image segmentation, which is usually extremely difficult to obtain using currently available computer-based algorithms (Figure 1). Moreover, the shape matching process based on contours can not be performed efficiently.

In our project, the adverse effect of inaccurate image segmentation is reduced by using a robust region matching metric, namely, the Integrated Region Matching (IRM) measure.

BACKGROUND

Content-based image retrieval systems roughly fall into three categories depending on the signature extraction approach used: histogram, color layout, and region-based search. There are also systems that combine retrieval results from individual algorithms by a weighted sum matching metric [5], or other merging schemes.

Region-based retrieval systems represent images at the object-level. A region-based retrieval system applies image segmentation to decompose an image into regions, which correspond to objects if the decomposition is ideal. Since the retrieval system has identified objects in the image, it is easier for the system to recognize similar objects at different locations and with different orientations and sizes. Region-based retrieval systems include the NeTra system [7] and the Blobworld system [1].

The NeTra and the Blobworld systems compare images based on individual regions. Although querying based on a limited number of regions is allowed, the query is performed by merging single-region query results. Because of the great difficulty of achieving accurate segmentation, systems in [7, 1] tend to partition one object into several regions with none of them being representative for the object, especially for images without distinctive objects and scenes. Consequently, it is often difficult for users to determine which regions and features should be used for retrieval. Not much attention has been paid to developing similarity measures that combine information from all of the regions.

Li, Wang and Wiederhold [6] developed a new metric, the IRM (Integrated Region Matching), for matching region features within a CBIR system. Experiments has shown much improved accuracy with fast speed. It is naturally suited for our pathology image retrieval purpose. Our pathology image retrieval algorithm can be combined with our previously developed wavelet-based progressive pathology image transmission and browsing algorithm [13]. The system is expandable for medical image databases.

METHODS

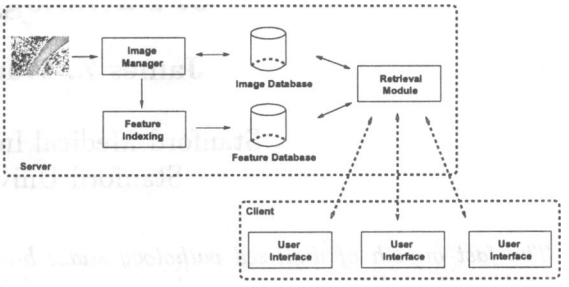

Figure 2: Basic structure of the system.

Our image retrieval system is a Web-based server-client system. It has several components, including an image database manager, a wavelet-based feature indexing module, a user query processing module, and a user interface. Figure 2 shows the basic structure of the system.

Image Database Manager

The image database manager manages the image files so that the feature-based image indexing module can efficiently process the images to extract features. This module can be combined with our previously developed wavelet-based progressive pathology image transmission and browsing algorithm [13] to efficiently provide the client with portions of the high resolution original image at a resolution specified by the client.

Wavelets-based Feature Indexing

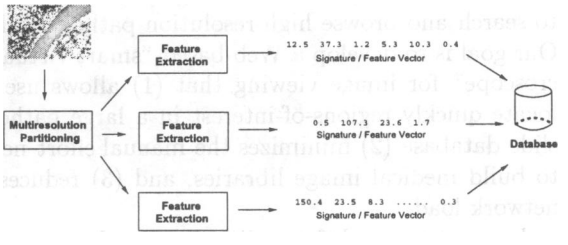

Figure 3: Structure of the image indexing system.

Wavelets are basis functions that have some similarities to both splines and Fourier series [4, 8]. They have advantages when the aperiodic signal contains many discontinuities or sharp changes. It decompose signals into different frequency components and analyze each component with a resolution matching its scale. Applications of wavelets [10, 12, 13] to signal denoising, image compression, image smoothing, fractal analysis and turbulence characterization are active research topics.

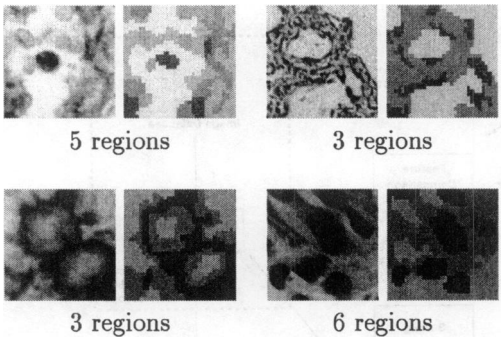

Figure 4: Segmentation results obtained using an algorithm based on k-means. A region is defined as a collection of pixels.

We index the block segments of the original images in different scales (Figure 3). This is done by partitioning an original image into smaller overlapping blocks of pixels in different scale. For the purpose of this project, we use wavelet to characterize the local texture properties within pathology images. We do not use color features because pathology images have limited colors. An image segmentation algorithm based on the k-means statistical clustering method is applied to determine the object configurations within each image segment. A region is defined as a collection of pixels. It may not be a connected region. This is much more flexible than most existing region segmentation. Figure 4 shows the segmentation examples. The details of the features are reported in [14, 6].

User Query Processing

The Web-based query processing server is illustrated in Figure 5. Given a query image segment or a sketch, the Web-based query processing server process the query to extract the relevant image features. Then the feature is matched against the features stored in the feature database to find the similar image segments in the database. The IRM [6] metric is used to determine the similarity. The metric is robust with respect to color distortions, shape distortions, cropping, rotation, translation and scale changes.

Once the similar image segments are determined, the user may interact with the wavelet-based virtual microscope server to browse the portions of the original image at the scale specified by the user.

Client Program

With the virtual microscope project, we have implemented a Web-based user interface that allows the users to magnify any portion of the pathological images

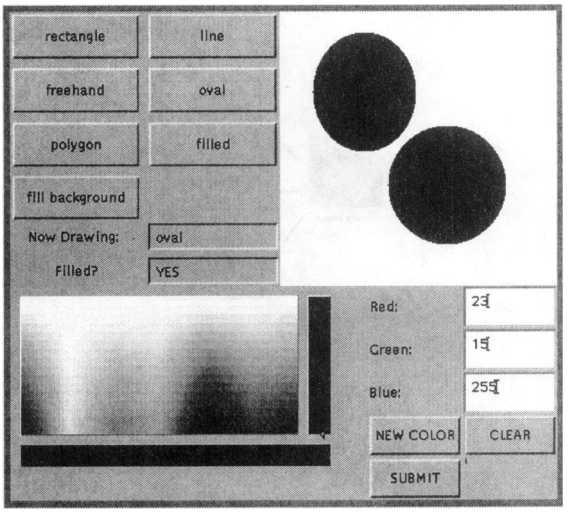

Figure 6: The JAVA user interface allows users to draw sketch queries.

in different level of resolutions. We use Web interface and JAVA primarily because the wide acceptance of the Internet and the Web in health care environments. The JAVA interface (Figure 6) allows the user to draw sketches with polygons, rectangles, ovals, and lines.

RESULTS

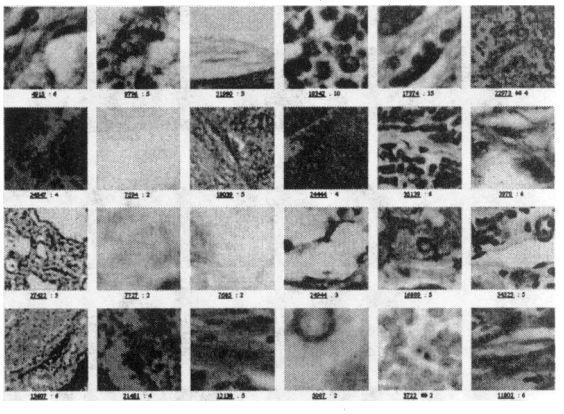

Figure 7: A set of randomly selected image fragments in the database.

The server program is implemented in C on a Pentium III 450 MHz PC running LINUX. It takes approximately two minutes of CPU time for the image database server and the feature extraction module to process a 24-bit color image of 2400×3600 pixels. This procedure is performed only once for each image in the database. The query processing is very fast. For a

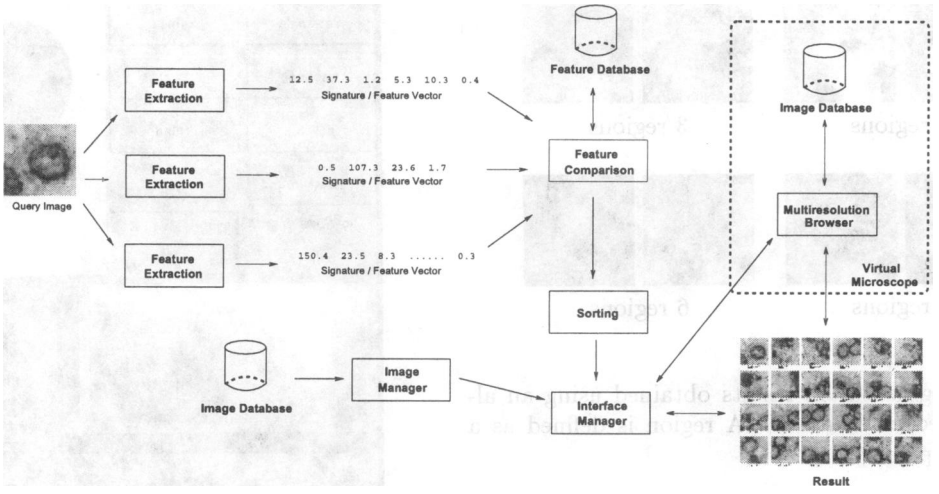

Figure 5: Structure of the Web-based query processing server.

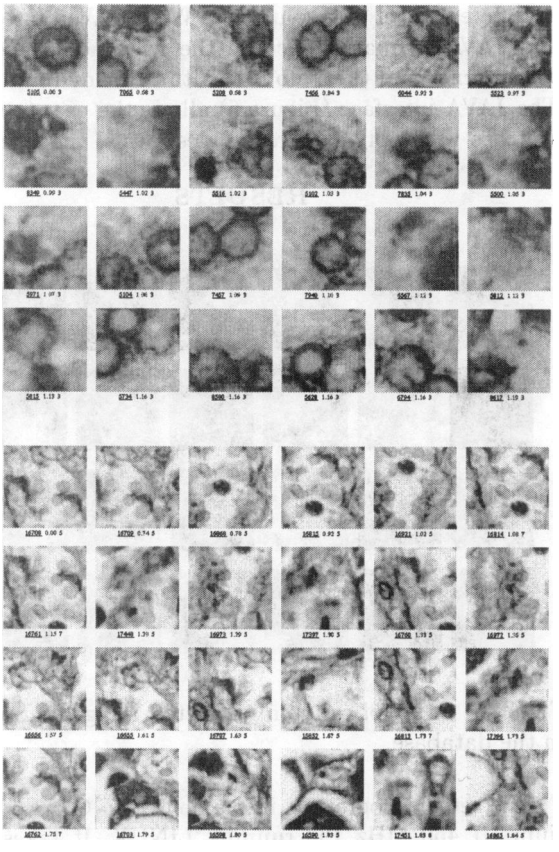

Figure 8: Sample image query results. The upper-left corner image in each block of images is the query. The images are listed based on the closeness to the query.

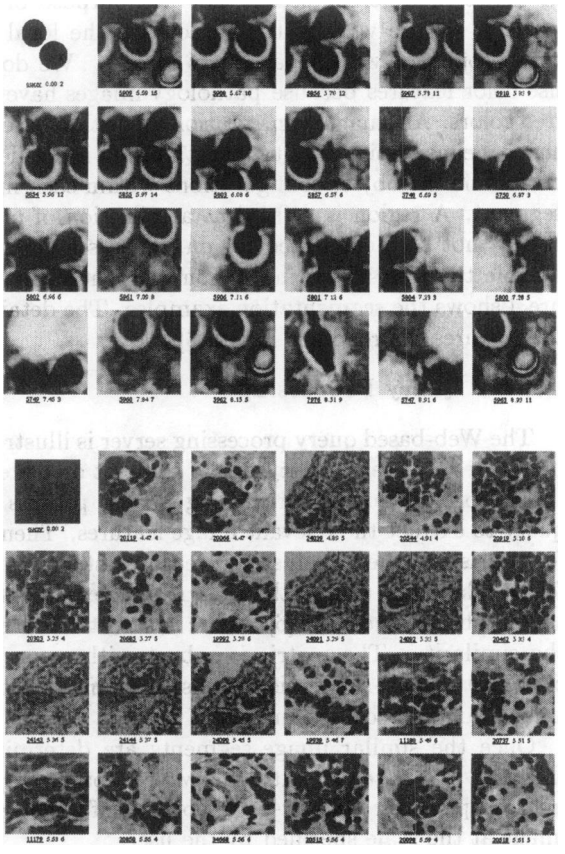

Figure 9: Query results for the drawing sketches. The upper-left corner image in each block of images is the query. The images are listed based on the closeness to the query.

database of 70,000 image fragments, it takes an average of about one second CPU time to retrieve the image segments similar to the one queried by the user. With our experiments, the system has demonstrated a precision of about 90% when examining the best 50 matches for each query. The relevance is defined by the similarity in the object configuration. Figure 8 shows sample image query results with the HTML-based client user interface. Figure 9 shows the query results for hand-drawing sketch queries.

CONCLUSIONS AND FUTURE WORK

In this paper, we have demonstrated an efficient region-based image retrieval system for pathology image databases. The system uses wavelets and IRM to index and match images. The system is practical for real-world applications. The algorithm can be combined with our previously developed wavelet-based progressive pathology image transmission and browsing algorithm. The system is expandable for medical image databases.

It is possible to improve the searching accuracy by further refining the indexing process. A better user drawing interface is also important to make the system more effective for the medical community.

Acknowledgments

We would like to acknowledge the valuable comments and suggestions from Jia Li of Xerox Palo Alto Research Center, Donald Regula of Stanford University, and Gio Wiederhold of Stanford University. We would also like to thank Desmond Chan and Xin Wang of Stanford University for their work on developing the JAVA drawing interface.

References

[1] C. Carson, M. Thomas, S. Belongie, J. M. Hellerstein, J. Malik, "Blobworld: A system for region-based image indexing and retrieval," *Third Int. Conf. on Visual Information Systems*, June 1999.

[2] D. Comaniciu, P. Meer, D. Foran, "Shape-based image indexing and retrieval for diagnostic pathology", Proc. 14th International Conf. on Pattern Recognition, Australia 16-20, Vol 1., pp. 902-4, Aug. 1998.

[3] C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic, W. Equitz, "Efficient and effective querying by image content," *Journal of Intelligent Information Systems: Integrating Artificial Intelligence and Database Technologies*, vol. 3, no. 3-4, pp. 231-62, July 1994.

[4] R. M. Gray, J. W. Goodman, *Fourier Transforms : An Introduction for Engineers*, Kluwer Academic Publishers, 1995.

[5] A. Gupta, R. Jain, "Visual information retrieval," *Comm. Assoc. Comp. Mach.*, vol. 40, no. 5, pp. 70-79, May 1997.

[6] J. Li, J. Z. Wang, G. Wiederhold, "IRM: Integrated Region Matching for image retrieval," *Proc. of ACM Multimedia*, Los Angeles, 2000.

[7] W. Y. Ma, B. Manjunath, "NaTra: a toolbox for navigating large image databases," *Proc. IEEE Int. Conf. Image Processing*, Santa Barbara, pp. 568-71, 1997.

[8] Y. Meyer, *Wavelets: Algorithms & Applications*, SIAM, Philadelphia, 1993.

[9] A. Pentland, R. W. Picard, S. Sclaroff, "Photobook: content-based manipulation of image databases," *SPIE Storage and Retrieval Image and Video Databases II*, San Jose, 1995.

[10] Special Issue on Wavelets and Signal Processing, *IEEE Trans. Signal Processing*, Vol.41, Dec. 1993.

[11] J. Z. Wang, G. Wiederhold, J. Li "Wavelet-based progressive transmission and security filtering for medical image distribution," *Advances in Biomedical Image Databases*, S. Wong (Ed.), Kluwer, 1999.

[12] J. Z. Wang, G. Wiederhold, O. Firschein, X. W. Sha, "Content-based image indexing and searching using Daubechies' wavelets," *International Journal of Digital Libraries(IJODL)*, 1(4):311-328, Springer-Verlag, 1998.

[13] J. Z. Wang, J. Nguyen, K.-K. Lo, C. Law, D. Regula, "Multiresolution browsing of pathology images using wavelets," *Proceedings of the 1999 American Medical Informatics Association (AMIA'99) Annual Fall Symposium*, 340-344, Washington, D.C., November, 1999.

[14] J. Z. Wang, J. Li, D. Chan, G. Wiederhold, "Semantics-sensitive retrieval for digital picture libraries", *D-LIB Magazine*, 5(11), November 1999. http://www.dlib.org